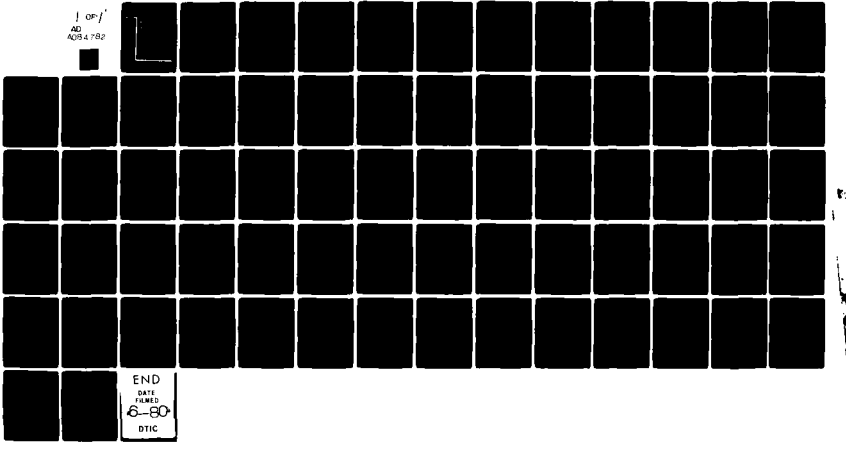
AD-A084 782 WESTINGHOUSE ELECTRIC CORP HUNT VALLEY MD F/G 5/9
OPERATIONAL CONSEQUENCES OF LITERACY GAP.(U)
MAY 80 J D KNIFFIN, C R STEVENSON, G R KLARE F33615-77-C-0048
UNCLASSIFIED AFHRL-TR-79-22 NL



AFHRL-TR-79-22
Supersedes AFHRL-TR-79-22, November 1979

AIR FORCE HUMAN RESOURCES LABORATORY

OPERATIONAL CONSEQUENCES OF LITERACY GAP

By

J. Douglas Kniffin
Calvin R. Stevenson
Westinghouse Electric Corporation
1111 Schilling Road
Hunt Valley, Maryland 21030

George R. Klare
Eileen B. Entin
Ohio University
42 Elmwood Place
Athens, Ohio 45701

Sharon L. Slaughter, Capt, USAF
Lydia Hooke

TECHNICAL TRAINING DIVISION
Lowry Air Force Base, Colorado 80230

May 1980

Final Report

AIR FORCE SYSTEMS COMMAND
BROOKS AIR FORCE BASE, TEXAS 78235

NOTICE

When U.S. Government drawings, specifications, or other data are used for any purpose other than a definitely related Government procurement operation, the Government thereby incurs no responsibility nor any obligation whatsoever, and the fact that the Government may have formulated, furnished, or in any way supplied the said drawings, specifications, or other data is not to be regarded by implication or otherwise, as in any manner licensing the holder or any other person or corporation, or conveying any rights or permission to manufacture, use, or sell any patented invention that may in any way be related thereto.

This final report was submitted by McDonnell Douglas Astronautics Company - St. Louis, P.O. Box 516, St. Louis, Missouri 63166, under contract F3361 5-78-C-0022, project 1121, with Technical Training Division, Air Force Human Resources Laboratory (AFSC), Lowry Air Force Base, Colorado 80230. Mr. Willi Nimms (TTO) was the Contract Monitor for the Laboratory.

This report has been reviewed by the Information Office (OI) and is releasable to the National Technical Information Service (NTIS). At NTIS, it will be available to the general public, including foreign nations.

This technical report has been reviewed and is approved for publication.

MARTY R. ROCKWAY, Technical Director
Technical Training Division

RONALD W. TERRY, Colonel, USAF
Commander

ADDENDUM

Kniffin, J.D., Stevenson, C.R., Klare, G.R., Entin, E.B., Slaughter, S.L., & Hooke, L. Operational Consequences of Literacy Gap. Lowry AFB, CO: Technical Training Division, Air Force Human Resources Laboratory, May 1980.

Due to printing error, change second paragraph of inside front cover to read as follows:

This final report was submitted by Westinghouse Electric Corporation, 1111 Schilling Road, Hunt Valley, Maryland 21030 under Contract F33615-77-C-0048, Project 1121, with the Technical Training Division, Air Force Human Resources Laboratory (AFSC), Lowry Air Force Base, Colorado 80230. Dr. Lydia Hooke was the Contract Monitor for the Laboratory.

E.L. ELLIOTT
Chief, Technical Editing

Unclassified

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE	READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER AFHRL-TR-79-22 Supersedes AFHRL-TR-79-22, November 1979	2. GOVT ACCESSION NO. AD-A084 782	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle) OPERATIONAL CONSEQUENCES OF LITERACY GAP	5. TYPE OF REPORT & PERIOD COVERED Final 6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s) J. Douglas Kniffin, Eileen B. Entin, Calvin R. Stevenson, Sharon L. Slaughter, George R. Klare, Lydia Hooke	8. CONTRACT OR GRANT NUMBER(s) F33615-77-C-0048
9. PERFORMING ORGANIZATION NAME AND ADDRESS Westinghouse Electric Corporation 1111 Schilling Road Hunt Valley, Maryland 21030	10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS 62205F 11210413
11. CONTROLLING OFFICE NAME AND ADDRESS HQ Air Force Human Resources Laboratory (AFSC) Brooks Air Force Base, Texas 78235	12. REPORT DATE May 80 13. NUMBER OF PAGES 64
14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office) Technical Training Division Air Force Human Resources Laboratory Lowry Air Force Base, Colorado 80230	15. SECURITY CLASS. (of this report) Unclassified 15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

literacy gap
preference measures
readability formulas
readability versions
reading comprehension tests
reading grade level (RGL)

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

Air Force managers and supervisors often face problems caused by reading difficulties among their personnel. These usually appear to be a joint function of level of reading skill of the personnel and level of difficulty of the materials they face. The term "literacy gap" refers to the difference between the two levels. This study proposed to examine the effects upon reading comprehension and preferences of three sizes of literacy gaps. This study involved three independent variables: Air Force personnel at tested reading grade levels (RGLs) of 8 and 10; Air Force job related materials, written at literacy gaps of 0, −2, and −4, and reading times of 30, 45, and 60 minutes. A gap of −2, for example, meant that the materials were two grade levels higher than the tested reading ability of the subjects. An additional question of interest was whether increasing the time allocated for reading would mitigate the

DD FORM 1 JAN 73 1473

Unclassified

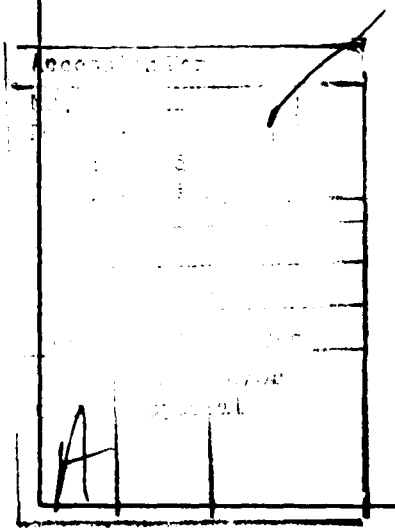
SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

Item 20 Continued :

effect of large literacy gaps. Thus, reading times of 30, 45, and 60 minutes were examined. Subjects were Air Force personnel at two levels of reading ability, 8th and 10th grades. Content was chosen from Air Force job-related material in two areas: Supervision and Safety/Sanitation. Reading materials consisted of 5250 word passages prepared at readability grade levels of 8, 10, 12, and 14, with content unchanged. Multiple choice tests of comprehension were also prepared and a short questionnaire designed to measure preferences was developed.

Preference measures indicated that readers judged lower-gap materials significantly easier to read and clearer than higher gap materials. When comprehension scores were analyzed, the results were as follows: (a) performance on the Safety/Sanitation passage was substantially better, (b) subjects with 10th grade reading ability performed consistently better than those with 8th grade ability, (c) greater literacy gaps led to poorer comprehension, but the effect of this variable was relatively small, and (d) longer reading times led to greater comprehension. However, comprehension did not increase in proportion to the amount of additional reading time; that is, a large amount of additional time invested in reading resulted in only a modest gain in comprehension. It should be noted that the effects, though statistically significant, were small and that the largest effect was due to subject matter rather than any of the variables of experimental interest.

The following recommendations for the Air Force were made: (a) before major efforts are undertaken to rewrite Air Force materials for greater ease of reading, it would seem expedient to determine how much a negative literacy gap influences actual job performance, (b) efforts to improve readability of materials should be directed at populations and situations where motivation and interest are unlikely to be high, and (c) while increasing reading time would seem to be a reliable and straightforward way to increase test comprehension, the results of this study indicated that the learning efficiency of this approach is not high. Therefore, in applying this approach to particular situations, it may be well to carefully analyze whether the gain in comprehension is worth the extra expenditure of reading time.



SUMMARY

OBJECTIVES

Air Force managers and supervisors often face problems caused by their personnel having reading difficulties. These problems appear to be a joint function of the level of the reading skill of the personnel and the level of the difficulty of the materials they must use. The term "literacy gap" refers to the difference between the two levels. A gap of -2, for example, indicates that a text is estimated to be written at a grade level two levels above that of its readers. This study proposed to measure the effects upon reading comprehension of three sizes of literacy gaps. An additional question investigated was whether increasing the time allocated for reading would overcome the detrimental effects of literacy gap.

APPROACH AND SPECIFICS

This study measured the effects of three experimental variables on comprehension of text passages developed from Air Force reading material. Subjects were tested with 5250 word passages which had been adapted from materials used in two Air Force career fields; the "Supervision" passage came from the Pavements Maintenance career ladder and the "Safety and Sanitation" passage from that of Diet Therapy Supervisor. The factors investigated were:

1. Reading ability: Air Force Personnel with identified reading grade levels of 8 and 10 were tested.

2. Literacy gap: 8th, 10th, 12th and 14th grade level versions of the two passages were developed. These versions were given to subjects at the two reading grade levels so as to create litercy gaps of 0, -2, and -4.

3. Reading time: periods of 30, 45, and 60 minutes were used, with testing occurring after every 15 minutes of reading.

All personnel read passages of the same length, but each person read only one of the two passages. Comprehension was measured by correctness of answers to a 52-item multiple-choice test. Personnel were subsequently asked to compare two versions of the passage they had not read previously in terms of readability, clarity, interest and information content.

RESULTS

All factors (subject matter of passages, reading ability, literacy gap and reading time) were found to affect scores on the comprehension tests at the .05 level of significance. The following are the results for each of the factors, averaged over levels of the remaining factors:

1. Comprehension on the Supervision passages was 61.7% correct and on the Safety passage was 78.6% correct.

2. Subjects at reading grade levels of 8 and 10 had scores of 67.4% and 71.6% respectively.

3. Literacy gaps of -4, -2, and 0 yielded scores of 67.4%, 69.7%, and 72.7% respectively.

4. Reading times of 30, 45, and 60 minutes yielded comprehension scores of 65.8%, 70.6%, and 73% respectively. However, comprehension did not increase in proportion to the amount of additional reading time.

It will be noted that effects, though significant, were small and that the largest effect was due to subject matter rather than the variables of experimental interest. Analysis of preference questions showed that 44% of the subjects failed to judge passages written at different levels to differ in readability or clarity. When subjects did judge that the passages differed, significantly more of them judged the passage written at the lower grade level to be clearer and more readable. Passages differing in grade level did not differ in judgments of interest or information content.

CONCLUSIONS

The literacy gap produced a small but significant effect upon comprehension scores under the conditions of this study, i.e., with relatively long passages of approximately 5000 words. One possibility suggested by previous readability research is that repeated testing during the experiment induced a high level of motivation in the subjects and that the liberal reading and testing times allowed this motivation to reduce the effect of text difficulty upon comprehension scores. Perhaps, too, the scarcity of appropriate subjects at the lower reading levels contributed to the attenuation. Increasing the reading time, for the range of times used here, appears to increase the text comprehension scores of readers. However, the relation between reading time and comprehension scores is such that subjects given more time learn less efficiently (i.e., learn less per unit time). The effect of added reading time appears to remain constant at all levels of literacy gap.

RECOMMENDATIONS

This study has resulted in the following recommendations:

1. Before major efforts are undertaken to rewrite Air Force materials for greater ease of reading, it would seem expedient to determine the extent to which a negative literacy gap influences performance on the job.

2. It is suggested that efforts to improve readability of materials might best be directed at populations and situations where motivation and interest are unlikely to be high.

3. Increasing reading time would seem to be a reliable and straightforward way to increase text comprehension. However, because of the decreased learning efficiency that this method is likely to induce, a careful analysis of whether the gain in comprehension is worth the extra expenditure of reading time should first be performed.

CONTENTS

LIST OF TABLES

LIST OF FIGURES

INTRODUCTION

For the last two decades, managers and supervisors in the Air Force have had to contend with special problems caused by reading difficulties among the airmen under their command. These so-called "literacy problems" can slow training schedules, lower performance on the job, and increase personnel costs. Studies directed at solving these problems typically fall under two distinct types of research and development efforts. The first approach, remedial training for personnel, sets up literacy training programs for the most seriously handicapped readers. The second approach, study of materials, examines the difficulty level of the reading matter and experimentally modifies that level where necessary. The research described in this report follows the second approach.

The strategy most often followed uses "readability formulas" to analyze the materials. These formulas typically provide ratings along a scale of difficulty which parallels the school grade scale. Burkett (1975) and Klare (1963; 1974-1975) provide summaries of the extensive readability literature in the military and civilian research communities.

The research in readability shows that written material often fails to match the skill level of intended readers, thus creating a "literacy gap' of one or more grades. Mockovak (1974a and 1974b) made an intensive study of available methodologies and then applied the most appropriate readability formula in an extensive examination of 56 Air Force career ladders. He found that 43, or almost 80%, had a "negative" gap, meaning that the readability grade level of the material exceeded the estimated reading grade level (RGL) of the intended readers. Of these, 29 had a gap greater than one, 17 greater than two and four had a gap greater than three grade levels.

The results of Mockovak's work indicate that the reading abilities of Air Force personnel and the reading demands of Air Force materials vary greatly across career ladders. Furthermore, significant gaps appear to exist between the reading skill levels of individuals and the reading requirements of their materials, even materials written at relatively low average difficulty levels. This situation typically occurs in those career ladders where lower aptitude levels suffice for entry. DeGuelle (1975) suggests that RGL estimates of those personnel with generally inadequate reading skills may themselves be low. Such work, and particulary the overview of Sticht (1975), reinforces the suggestion that literacy gaps can create potential problems for Air Force training and operational efficiency.

Klare (1969) labelled research of the above sort "prediction" research, since readability formulas can "predict" the reading difficulty level of materials. "Production" research, on the other hand, involves a test of whether modifying the readability of the material will actually make it more appropriate for intended readers. A subsequent article (Klare, 1976) compares the two research approaches and suggests the problems likely to be found with each. Findings based on an intensive analysis of 36 experimental studies suggest that a number of variables may affect the likelihood of significant results in production research. Of the 36 studies, 19 showed that making writing more readable produced a significant increase in comprehension, and 11 showed that it did not. Six of the studies produced mixed results - some differences were significant and some were not. Detailed analysis covered 28 characteristics in each study, grouped under the following general categories:

1. The experimental passages and how they were modified.
2. The tests and other dependent measures used.
3. Descriptions of the subjects and their characteristics.
4. The instructions given to the subjects.
5. Details of the experimental situation.
6. The statistical analysis employed.
7. The results and the detailed discussion based on them.

Such expected variables as quality of the rewriting or of the test used appeared to affect the probability of observing non-significant results in certain cases. However, the chief factor--surprisingly--appeared to be reader motivation Two interacting aspects appeared responsible:

1. Conditions which raised the level of reader motivation (e.g., promised reward or threat, or the experimental situation itself), in combination with

2. Conditions which allowed the increased level of motivation to reduce an effect (e.g., liberal time for reading and/or testing time).

The review study (Klare, 1976) suggested a model of the variables in the experimental situation likely to affect comprehension scores when readability has been modified. Three recent studies have supported the predictions from the model. Denbow (1973) found that improved readability produced significant information gain with each of two passages of different content. The amount of gain

attributable to readability was however, significantly greater with the non-preferred content, as the model predicts. Fass and Schumacher (in press) showed that, in a similar fashion, monetary reward coupled with liberal reading/testing time could wipe out the demonstrated effects of readability upon comprehension. Entin and Klare (in press) showed that correcting multiple-choice comprehension scores for subjects' "prior knowledge" of passage content increased the correlation with readability scores on the passages.

Production experiments covered in the above review study, (Klare, 1976), though rather sizable in number, require further examination and refinement. The skill level of readers, for example, might be measured more carefully (and not estimated), in order to achieve greater precision in specifying literacy gap. Materials might be prepared at a number of readability levels, so that the gaps themsleves can be varied and relative effects compared. Reading and testing time might be varied to observe the effects upon comprehension scores. And, of course, subjects having low-ability might be used, since their deficiencies are most likely to have an impact on Air Force training and operational efficiency.

The above review study (Klare, 1976) suggested a further addition: that Air Force personnel be asked to indicate prefferences among the several readability levels. Even where modified readability failed to produce significant differences in comprehension, reader preferences generally favored the more readable versions. Consequently, readers in the present study were asked to compare samples of writing at different levels and make preferential judgments.

Finally, the literature in the area of comprehending and/or learning from prose indicates the desirability of using more than one type of content in research. Findings may otherwise be content-specific and may not generalize to other contents. Denbow (1973) suggests that the research include two contents differing in preferability, since this variable affected his experimental results. And, since Mockovak (1974a and 1974b) had done a considerable amount of research on Air Force job related materials, selecting experimental passages from among such materials seemed highly desirable and feasible.

The background research and the objectives described above led to plans for a 3 x 3 x 3 factorial design. The intended factors included the following:

1. Subjects at three reading grade levels of 6, 8, and 10.

2. Three literacy gaps 0, -2, and -4 for each reading grade level, meaning that experimental passages at readability grade levels of 6, 8, 10, 12 and 14 were needed.

3. Three time periods (30, 45, and 60 minutes), with each period divided into 15-minute segments for reading and followed by testing over the material read.

At this point note should be taken that reading time and number of testings co-vary in this design. Consequently, the effects of reading time and testing are confounded. Put another way, either reading time or amount covered in a test (or both) could affect comprehension performance as measured by a multiple-choice test. Although this was recognized at the time, an alternative testing procedure, e.g., one test covering all the material at the end of the session, would have had an even more serious consequences. In that case, it was felt that test performance would have reflected memory factors more than comprehension. Since access to subjects and total testing time were constrained by operational needs of the Air Force, a less than optimum design was deemed acceptable.

Two sets of Air Force job related materials, one on Supervision and one on Safety and Sanitation, provided the passages for experimentation.

The original plans called for subjects to indicate their preference for one of two readability versions by making judgments on approximately 2500-word segments (i.e., approximately one half) of the content they had not read for comprehension. The demands this created for testing time, however, required that these preference passages be drastically reduced in size; therefore, 200-word segments were substituted. The need to eliminate any possible judgment differences owing to content or order of presentation rather than readablity (the desired variable) led to a counter-balanced design for the two experimental versions. Details of the designs actually used and other aspects of the research are presented in the next section.

METHOD

This section is divided into the following sub-sections:

1. Subjects
2. Materials
    a. Reading Test
    b. Experimental Written Materials
    c. Comprehension Tests
    d. Preference Measures
3. Experimental Designs
    a. Comprehension Testing
    b. Preference Measurement
4. Procedure

Subjects

Air Force needs, as mentioned earlier, dictated the use of personnel with reading abilities at the 6th, 8th, and 10th RGLs. Experimental considerations further required selecting the subjects within a narrow range around each of these grade levels, the choice being those within a 95% confidence interval around each. Finally, typical measurement of reading comprehension with multiple-choice items suggested a minimum of 7, and preferably 10, subjects as desirable for each cell of the experimental design.

During the period of experimental testing, the average RGL of the personnel being tested was approximately 11, so very few lower grade level subjects became available. This necessitated modifying each of the ideal requirements as indicated below.

1. The original plan to obtain all experimental subjects from basic trainee flights at Lackland AFB could not, it soon appeared, provide personnel at the 6th RGL and probably not enough at the 8th RGL. A discussion with Air Force Human Resources Laboratory personnel suggested the following:

    a. Extending the experimental testing period long enough to test at least 90 subjects having an 8 RGL.

    b. Modifying the experimental design to include two, rather than the intended three, reading grade levels. The decision to reduce the levels to be analyzed was necessitated by the fact that a total of only 11 subjects at the 6th RGL had been tested by the end of the extended test period. Another possibility, use of subjects at the 12th

RGL, seemed undesirable, since personnel with reading abilities at that level create few problems for the Air Force. Furthermore, a complete factorial design using such subjects would have required additional rewriting of experimental materials in order to have the planned number of literacy gaps for them. Such a change would have meant an unacceptable delay in the experimentation schedule.

2. The original plan to select for experimentation only those personnel within a 95% confidence interval around each grade level could not be carried out due to the lack of sufficient subjects within these intervals. Since testing at other locations in the San Antonio area could not have provided enough additional subjects, the following steps were taken.

a. The confidence interval for the 10th RGL was widened to 99.9% to avoid eliminating a number of subjects. This interval, though broader, departed relatively little from the intended 95% interval. (Details of the intervals are provided in the following sub-section, "Reading

b. The confidence interval for the 8th RGL was widened to 99.99999%. With a 95% interval, many subjects would have had to be eliminated and within-cell numbers would have been totally inadequate. The experimental design could still be carried out with little disruption but with some loss of precision in analysis involving 8th RGL subjects.

3. The original plan to test a minimum number of 7, and preferably 10 subjects per cell, could not be carried out. The actual number of subjects available led to the following modifications in the original design.

a. At the 10th RGL, a total of 143 subjects were tested. This turned out to be no fewer than 7 nor more than 9 subjects per cell. The mean value of 7.94 subjects per cell was close to the desired figure.

b. At the 8th RGL, 97 subjects were tested. This translated into as few as three subjects in one cell, four in three other cells, and seven in only three cells. The mean value of 5.39 per cell necessarily resulted in some loss of precision for analyses involving 8th RGL subjects.

Materials

Reading Test. Establishing an adequate "literacy gap" required that the RGL of experimental subjects be determined precisely. The common practice of using "last school grade completed" cannot satisfy this requirement, whereas a reading test can. Examination of a number of reading tests

indicated that the California Achievement Tests (Tiegs & Clarke, 1970): Reading, Level 4 (Grades 6-9) came closest to meeting this need. The actual testing involved administering the Vocabulary and Comprehension portions of the California Achievement Test (Tiegs & Clarke, 1970), Form A, 1970 Edition. Although the grade span (6-9) did not include Grade 10, Level 4 was selected for the following reasons:

1. Level 4 was used by the Air Force where preliminary screening of personnel suggested the need for more intensive testing of reading comprehension of personnel.

2. Norms were available for total reading scores (vocabulary plus comprehension) at grade equivalents of levels of 0.6 to 13.6.

3. According to the examiner's manual (Tiegs and Clarke, 1970), the time limits are so constructed that below-average students in the lowest grade of the grade span of a level have ample time to attempt every item.

The standard deviation and number of cases needed to compute standard error of the mean were 15.83 and 383, and came from the norms tables in the manual. The 95%, 99.9%, and 99.99999% confidence intervals based on these values are presented in Table 1.

Table 1

The 95%, 99.9%, and 99.99999% Confidence Intervals for the 6th, 8th, and 10 RGL's based on the California Achievement Tests (Tiegs & Clarke, 1970): Reading, Level 4 (Grades 6-9), 1970 Edition.

Confidence Intervals	Raw Scores by RGL 6	8	10
95%	35-38	50-54	62-65
99.9%	33-40	49-55	60-67
99.99999%	31-42	48-57	58-69

The total reading scores (raw scores) of the individual subjects who were tested, presented in Table 2, show that with the exclusion of the first five cases (scores below 48):

1. The 97 subjects at the 8th RGL fall within the 99.99999% confidence interval, 48-57; and,

2. The 143 subjects at the 10th RGL fall within the 99.9% confidence interval, 60-67.

The first case could not be used because the reading score could not be specified. The next four cases could not be used because, though they might have been considered part of the 8th RGL group in an emergency, they actually fell between the 6th and 8th RGL groups. All other subjects became part of the analysis.

Table 2

Frequency Distribution of Total Reading Scores (Sum of Vocabulary and Comphrension Raw Scores) on the California Achievement Tests: Reading, Level 4 (Grade 6-9), 1970 Edition, and Mean Values for Experimental Subjects.

Frequency Distribution		Mean Values for Subjects Used by Grade Levels and Contents	
Total Reading Raw Scores	Number of Subjects		
--	1		
44	2 *		
46	2 *		
48	1		
49	6		
50	12	Grade Level	Mean
51	13	8(N=97)	53.2
52	6	10(N=143)	62.3
53	9		
54	13		
55	14		
56	15		
57	8		
60	4	Content	Mean
61	25	Supervison(N=117)	59.5
62	14	Safety & (N=123)	59.4
63	20	Sanitation	
64	18		
65	31		
66	31		
	245		

* Scores below 48 were not included in the analysis.

Experimental Written Materials. As noted earlier, Mockovak (1974a & 1974b), showed that Air Force career materials vary greatly in readability level, and generally fall beyond the estimated reading skill level of intended readers. Such materials become prime prospects for experimentation, since experimental results might well come to have direct and widespread practical consequences for the Air Force. The large variety of job related materials, furthermore, offered excellent opportunities for the selection of experimental materials.

Passages of approximately 5,250 words came from each of two Air Force Career Development Courses (CDCs). The first passage, referred to as "Supervision," is found in CDC 55150, Pavements Maintenance Specialist, Volume I, pages 30 to 36. The second passage, referred to as "Safety and Sanitation", is found in CDC 62271, Diet Therapy Supervisor, Volume I, pages 55 to 66. The length of the passages was determined by the following considerations.

1. The Air Force specified a normal reading rate of 175 words per minute for the subjects with a minimum reading time of 30 minutes. This meant approximately 5,000-word passages were needed.

2. The requirement for reading periods of 30, 45, and 60 minutes dictated convenient division into halves, thirds, and quarters. Thus, the number 5,250 became a desirable figure.

The particular passages selected met, in addition, the following requirements.

1. Freedom from large numbers of illustrations or tables integrated with the text, since their presence would have made experimental rewriting and analysis difficult.

2. Freedom from large groups of numbers and acronyms, since these also would have made experimental rewriting and analysis difficult.

3. Readability as close as possible to 10th grade level. This meant that the rewritten versions of the passage could be "written up" and "written down" to about the same degree.

A further characteristic concerned the preference-value of the materials. As noted earlier, Denbow (1973) showed that readability made less difference with high-preference than with low-preference material. Consequently, materials of low and middle preference appeared desirable. Air Force personnel familiar with subject preferences identified Supervision materials as low-preference and Safety and

Sanitation materials as middle-preference contents.

The literacy gaps selected for study, 0, -2, and -4 (grade level gaps), required preparing versions of each content at readability grade levels of 6, 8, 10, 12, and 14. Thus, subjects with tested RGLs of 6 could read passages at 6, 8, and 10 readability levels. Those at RGL 8 could read at levels 8, 10, and 12, and those at RGL 10 could read at levels 10, 12, and 14. The unavailability of subjects at RGL 6 made the two readability versions at level 6 unusable, but the other versions were used as intended.

Preparation of the readability versions followed the steps outlined below:

1. Precise word counts of the original versions of both experimental passages were made. This was done to assure that the passages would properly divide into halves, thirds, and quarters. Minor changes, usually deletions from the original text, were made where possible to obtain the desired division points.

2. Each experimental passage was then split into several consecutive shorter sections of about 200 words each. Accurate word counts were made on each short section, and care was taken to assure that each of the short sections addressed only one main topic. The Supervision passage, divided into 26 short sections, and the Safety and Sanitation passage, divided into 27 short sections. The division of the experimental materials into short sections was done for the following reasons:

a. Readability versions of each passage at grade levels 6, 8, 10, 12, and 14 were needed. The best way to assure that these target grade levels would be met was to make the writing within each version as consistently close to the target grade level as possible. Working with small units of text greatly facilitated the production of rewritten text that was consistently near a target grade level.

b. Readability formula calculations on complete passages, especially long passages, can be somewhat misleading. This is because an average readability level does not fully reflect the range of difficulty of selected sections of the passage. Because so many readability versions of the experimental passages had to be prepared, it was necessary to verify precisely the grade level difficulty of all sections of the original passages.

3. Readability grade level calculations on each of the original CDC passages were performed. Individual calculations were performed on all the short sections with

the Supervision and Safety and Sanitation materials. These calculations provided the data on the degree to which each of the short sections had to be "written up" or "written down" to meet the target grade level of the various readability versions. The readability grade level of the CDC passages was determined using the Kincaid version of the Flesch Reading Ease formula (Kincaid, Fishburne, Rogers and Chissom, 1975). The formula is:

$$\text{Grade Level} = .39(\text{words/sentence}) + 11.8(\text{syllables/word}) - 15.59.$$

The Kincaid formula was most appropriate for several reasons.

a. The formula was developed using passages from military training materials and using military enlistees as subjects.

b. The formula scores are expressed as reading grade level equivalents.

c. The formula was developed on materials ranging in difficulty from about the 5th through the 16th grade levels. The formula thus provided accurate scores for all of the experimental materials developed under this program.

4. The 10th grade readability versions of both Supervision and Safety and Sanitation were prepared first. This was necessary because the item analysis tryout of the Comprehension Test items was to be based on 10th grade level materials and subjects.

Production of the 10th grade level readability versions was accomplished as follows for both the Supervision and Safety and Sanitation contents:

a. The readability formula data for each short section was analyzed. If the original text of a short section was above the 10th grade level, the text was rewritten to make it more readable. Conversely, if the original text of a short section was below the 10th grade level, the text was rewritten to make it less readable. Text was made more readable by following the suggestions outlined in A Manual for Readable Writing (Klare, 1975). Conversely, text was made less readable by using the reverse of the suggestions in Klare's manual. Klare's suggestions for making materials more readable consist of those changes in word and sentence variables which have a research basis. Without elaboration, the word changes are:

(1) Use familiar or frequently occurring words.
(2) Use short words instead of long words.
(3) Use words with high association value.

(4) Use concrete words instead of abstract words.
(5) Use active verbs instead of nominalizations.
(6) Limit or clarify the use of pronouns and other anaphora.

The sentence changes are:

(1) Write short sentences and clauses.
(2) Form statements instead of questions where possible.
(3) Make positive instead of negative statements where possible.
(4) Make statements in active instead of passive voice where possible.
(5) Change or avoid self-embedded sentences.
(6) Change constructions that are high in word depth to ones that are low.

The total number of words in all readability versions had to remain nearly constant. Therefore, the number of words in each rewritten short section was kept as close as possible to the number of words in each original short section. Each short section was rewritten up or down as necessary, without regard to formula score. The formula was then applied to the rewritten short sections to determine if it was near the 10th grade target level. If the formula score was close to the target level, then work on the next short section was started. If the target level was not met the short section was again rewritten and the formula applied again. This was repeated as often as necessary on each short section until the target grade level was met.

b. The manuscripts of the 10th grade versions of the Supervision and the Safety and Sanitation contents were then submitted to a panel of five "technical experts". Each individual expert was asked to compare the original CDC texts with the rewritten 10th grade manuscripts to determine if the meaning of any portion of the original CDC was was changed during the rewriting process. The experts, lead technical writers and editors of the Technical Logistics Data Department of Westinghouse, prepared their comments concerning changes in meaning. The comments were collected and changes were made to the manuscripts as needed to assure that there were no content differences between the original CDC's and the rewritten manuscripts.

c. The final overall reading grade level and length in words of 10th grade manuscript were then calculated. The readability formula score for Supervision was 10.0 and the length was 5251 words. The formula score for Safety and Sanitation was also 10.0 and the length was 5240 words.

5. Readability versions of the Supervision and Safety and Sanitation contents were then prepared at grade levels 6, 8, 12, and 14. The same process was used to prepare these versions as was used to prepare the 10th grade versions. The short sections within each of the original CDC materials were "written up" or "written down" as necessary, without regard to formula, until the target grade level was met. And, of course, efforts were again made to keep the length of each readability version the same as the original materials. Each readability version was submitted to a panel of five technical experts to determine that the meaning of the original material was not changed during the rewriting process. Again, comments of the experts were incorporated as necessary to assure that there were no content differences between the original CDCs and the rewritten versions.

The final overall reading grade level and length in words of each manuscript were then calculated. For the final readability versions of the Supervision content, the readability formula scores and word lengths were as follows: 6th grade version, formula score 5.9, length 5249 words; 8th grade version, formula score 8.0, length 5249 words; 12th grade version, formula score 11.9, length 5251 words; and 14th grade version, formula score 13.9, length 5246 words. For the final readability versions of the Safety and Sanitation content, the readability formula scores and word lengths were as follows: 6th grade version, formula score 6.0, length 5240 words; 8th grade version, formula score 8.0, length 5240 words; 12th grade version, formula score 12.0, length 5240 words; and 14th grade version, formula score 13.9, length 5241 words.

6. Final printed copies of all readability versions of both the Supervision and Safety and Sanitation contents were prepared. All versions of sample paragraphs from each content are given in the Appendix. For each content and for each readability version, there were three sets of experimental passages prepared. One set was split into halves for use by subjects who would be allowed one-half hour of total reading time during comprehension testing. A second set was split into thirds for use by subjects who would be allowed 45 minutes of total reading time during comprehension testing. The third set was split into quarters for use by subjects who would be allowed one hour of total reading time during comprehension testing.

The materials for the preference measure were extracted from the various readability versions of the Supervision and of the Safety and Sanitation contents. This meant that subjects who were tested for comprehension on a Supervision content were asked to give preference judgments on materials extracted from Safety and Sanitation and vice versa. The

criteria for selecting the particular short passages used in the preference measure were as follows:

a. The number of words in each half of the preference measure materials was virtually identical. This was done to avoid any possible preference bias toward a shorter or longer passage.

b. The first and second halves of the preference materials (each half at a different readability grade level) were selected from one continuous section of text extracted from the original materials; care was taken to assure that the general subject of each half was the same. This was done to avoid any possible bias toward one subject matter as opposed to another and to provide continuity between the first and second halves.

c. The first and second halves of the preference materials contained the same number of paragraphs. This was done so the first and second halves would have a similar appearance in print and so the content of one of the halves did not appear to be more formidable than the other.

d. An attempt based strictly on judgment, was made to assure that the first and second halves were equal in information content. This was done because one of the questions on the preference measure related to information gain.

Further information explaining the rationale for taking preference measure data is provided in the Preference Measure sub-section of this section.

Comprehension Tests. The original Air Force research requirements suggested either a multiple-choice comprehension test or a CLOZE comprehension test. Comparison of the two for the purposes of this study showed advantages for the multiple-choice test. These centered around the following:

1. The length of typical CLOZE tests, which have a 1:5 deletion ratio, prohibited their use in this experiment. Approximately 1,050 items would have been required for such a test, demanding an inordinate amount of subject time. Even a modified (shortened) CLOZE test of sufficient length to be satisfactory would have taken a great deal of time to answer.

2. Item analysis procedures fit traditional multiple-choice tests better than CLOZE tests.

3. Multiple-choice tests appeared more realistic for this experiment, since the Air Force uses them more

generally for measurement. CLOZE tests do have advantages in certain situations (see Klare, Sinaiko, Stolurow, 1972), notably convenience of construction and scoring, and closer relationship to readability measures (see Miller, 1972), which would have increased the chances of reliable results in this experiment. But the advantages of the multiple-choice method clearly outweighed the CLOZE advantages in this instance.

The development procedures for the multiple-choice tests included the following for each of the 5,250-word contents:

1. Writing 200 trial items, based upon the 10th RGL version of each content. This version appeared best for the purpose since it fell midway between the versions needed for experimentation (i.e., 6, 8, 10, 12, and 14). Each multiple-choice item contained a stem and four choices. The choices were so arranged that the correct choice would appear at each position an approximately equal number of times. To achieve this, the following random permutations were used:

DABC ADCB DACB DBCA
ABDC BACD BCDA CDAB
CBAD BDAC ABCD DBAC
DCAB BADC CADB BCAD
BDCA ADBC ACDB CDBC
CABD CBDA ACBD DCBA

The items were extracted from the text of each reading passage and were written under the following specifications:

a. The items should be in the same order as the text materials on which they are based.

b. The essence of the problem should be in the stem. Generally, the stem should be longer than any of the options, although there are exceptions (i.e., literature tests). Moreover, the stem must consist of a statement or question that contains a verb.

c. Repetition of key words in the options should be avoided.

d. The options should be listed below the stem in some order. Let the first option represent the correct option in the preliminary writing. The order of the options will be randomized later.

e. Responses or options should be plausible and homogeneous.

f. The correct answer should be no longer than the incorrect choices.

g. Irrelevant clues should be avoided.

h. The "all of the above" option should be avoided.

i. The "none of the above" option should be used sparingly. Moreover, if used, it should be the correct response about as many times as the incorrect one.

j. Four options per item should be sufficient, unless the maximum that can be written and still be plausible is only two or three.

k. Overlapping items should be avoided. For example:

(1) More than 150 pounds
(2) More than 160 pounds
(3) . . . .
(4) . . . .

If "2" is correct, then "1" is also correct.

l. The correct option should be completely correct or clearly adequate. Likewise, the incorrect options should be plausible, but thoroughy wrong or completely inadequate.

2. Running tryouts of 104 items selected on the basis of adequate coverage of the passages to basic trainees at Lackland AFB. The items were reduced from the original 200 to minimize the amount of time required. (Using the original 200 would have required 60 minutes of reading time plus 200 minutes of test time plus time to distribute materials and explain procedures, or between 4.5 and 5 hours).

Average syllables per word were determined for each item to obtain an overall readability grade level estimate for each of the two sets of 104 tryout items. Both sets were close to the 1.55 average syllables per word which is typical of 10th grade level text. Actual values for the Supervisior items were 6645 syllables/4306 words=1.54 and for the Safety and Sanitation items, 4848 syllables/3161 words = 1.53.

Tryout subjects read for four 15-minute periods and answered 26 items at the end of each reading period. The 60 minutes allowed for reading a passage and the 45 seconds allowed to respond to each item resulted in complete coverage of the materials by the tryout subjects.

Original plans to use tryout subjects at the 10th reading grade level could not be accomplished, since

TABLE 3

TRYOUT TEST STATISTICS FOR THE SUPERVISION AND SAFETY AND SANITATION CONTENTS

Title of Content	Female[1]	Male[2]	Total[3]	Mean	Standard Deviation	Reliability Coeff.	Error of Measmt.	Difficulty Mean	Difficulty Range	Discrimination Mean	Discrimination Range
Supervision	43	162	205	76.78	11.03	0.89	3.71	75.31%	19-98%	0.29	-0.01 to 0.63
Safety and Sanitation	29	204	233	89.18	9.11	0.88	3.13	85.71%	33-99%	0.38	-0.03 to 0.52

[1] Mean Madden-Tupes RGL = 12.00

[2] Mean Madden-Tupes RGL = 12.32

[3] Mean Madden-Tupes RGL = 12.24

determination of the basic trainees reading levels could not be estimated using the Madden-Tupes conversion (1966) prior to the actual tryouts. As Table 3 shows, of the 438 trainees tested, 205 (43 female and 162 male) were administered the tryout test for the Supervision passage, while 233 (29 female and 204 male) were administered the tryout test for Safety and Sanitation. Mean Madden-Tupes RGL calculated after the administration of the tryouts was 12.00 for female trainees, 12.32 for male trainees, and 12.24 for all trainees.

Also shown in Table 3 are performance data for the tryouts. In general, these data indicate that the tryout test for the Supervision content was more difficult than that for the Safety and Sanitation content. Scores for both content tests had similar reliabilities (and satisfactory indices). Mean difficulty percentages reflected the mean number of correct responses. On the average, there were no significant differences in item discrimination across the tests for the two contents.

3. Using traditional item-analysis procedures involving computation of the following:

a. Total percentage selecting the correct response to each item to provide difficulty index values as well as the percentage selecting each distractor.

b. Biserial correlations on the upper and lower 27 percent passing each item to provide discrimination index values.

These item-analysis procedures also yielded means, standard deviations, Kuder-Richardson Formula 20 reliability coefficients, standard errors of measurement, mean difficulty, and mean discrimination indices for the total test on each of the two passages.

4. Selecting the 52 items with the highest discrimination indices within the constraints of the following:

a. No discrimination index less than .20.

b. Percentage passing the item above chance level.

c. Divisibility such that sub-tests could be constructed with an appropriate number of items for the two, three, and four 15 minute reading periods, or 30, 45, and 60 minutes respectively (see the sub-section on Procedure for additional details).

d. Readability level, overall, of 10th grade.

5. Using the selected 10th reading level items and 6th, 8th, 12th, and 14th grade (content unchanged) readability passages to prepare items at each of these four other levels. This procedure was used to avoid introducing a bias of the test toward a particular level of reading material. The common criticism of tests, that "one can write easy items on difficult content or difficult items on easy content," may thus be less cogent.

Table 4 gives pertinent item-analysis data for the tests for each content: (a) Supervision and (b) Safety and Sanitation. Reliability tests could not easily have been run on the tests at this point, since only tryout data on a longer form could have been used. However, careful item selection afforded some assurance of reliability, with a reliability check planned on the experimental data itself (see the Results section for details).

Table 4

Pertinent Item Analysis Data for Multiple-Choice Tests for the Supervision and Safety and Sanitation Contents Written at the 10th Readability Grade Level

Title of Content	No. of Items	Difficulty Level Mean	Stand. Devia.	Range	Discrimination Index Mean	Stand. Devia.	Range
Supervision	52	74.21%	17.53	30-96%	0.36	0.08	.21-.63
Safety and Sanitation	52	87.63%	9.06	64-99%	0.37	0.09	.18*-.52

* Subdivision of content passage necessitated inclusion of one item with a discrimination index of 0.18. All other item discrimination indices were 0.20 or larger.

Preference Measures. Klare, in the examination of 36 studies attempting to increase reading comprehension by modifying readability (Klare, 1976), found that a number of variables could reduce the chances of significant results. He also found that even where significant increases in comprehension were not observed, subjects typically preferred the versions that were more readable to those which were less so. Subjects, furthermore, were able to make preference judgments relatively easily and reliably by reading somewhat briefer segments of the passages than those used in comprehension testing itself. In the present study, this procedure required that subjects getting one passage (content) for comprehension testing base their judgments on

comparisons of more versus less readable segments of another passage (content) to avoid interference. Also, since comparison could be made after comprehension testing, little disruption of normal activities or added experimental time needed to be introduced.

Consequently, preference measures were added in the experiment being described here simply by:

(a) asking a subject to read two passages of approximately 200 words each, one written at a more readable and one at a less readable level;

(b) having a subject who had read the Supervision content for comprehension purposes then read the Safety and Sanitation content for preference purposes, and vice versa; and

(c) scheduling the preference passages and questions after the comprehension testing had been completed.

The preference questions asked subjects to judge whether one passage, compared to the other seemed easier, more informative, more interesting, and clearer. Each of the four questions provided an opportunity for a subject to say he or she found no difference in the two passages. A fifth question, placed after the other four, asking whether subjects felt tired at the end of the experiment, was intended to allow investigation of possible fatigue effects.

The experimental design used in this part of the experiment can be found in the Preference Measurement sub-section of The Experimental Designs section which follows.

Experimental Designs

Comprehension Testing. The original plans called for a 3x3x3 factorial design for each content with the following three factors:

(a) subjects at three reading grade levels,6, 8, and 10;
(b) literacy gaps at three levels, 0, -2, and -4; and
(c) reading times at three levels, 30, 45, and 60 minutes.

As noted previously, sufficient subjects could not be found at the 6th RGL to use a 3 x 3 x 3 design. Sufficient numbers of subjects at the 8th and 10th RGLs were available to make possible a two-level subjects factor. The other two factors, literacy gaps and reading times, were used as planned which made a 2 x 3 x 3 design possible. The study may also be thought of as conforming to a single 2 x 2 x 3 x 3 design.

Preference Measurement. The original plans called for a counterbalanced design in which passages at two levels of readability were compared by subjects, who then stated their preference for one or the other. This arrangement provides a "benchmark" for comparison, since two passages can be compared directly with each other (see Klare, Mabry, Gustafson 1955, Frase, Schwartz, undated). However, both content and order can affect such preferences, so a counterbalanced design must be used to eliminate these effects. Though ratings on single passages might have been used to obtain preferences, such an arrangement would have produced less reliable data since it does not provide a "benchmark" for comparision. (Nor for that matter, does it handle content or order effects directly).

Preference judgments cannot usually be easily elicited from subjects without undesirably elaborate instructions regarding the bases for judgments. Consequently, four simple questions appeared desirable, with a provision in each for subjects to indicate they saw no differences between the passages. The probable loss of data with a no-difference option seemed to be preferable to the greater amount of, but more unreliable, data obtained with a forced-choice arrangement.

The four questions, as noted earlier, concerned judgment of which of a pair of passages seemed easier, more informative, more interesting, and clearer. Plans for analyzing the data from each question involved the design presented in Figure 1. Note that the literacy gaps form the basis for the comparisons of the paired passages and that the arbitrary specifications A & B, C & D, and E & F designate the comparisons made by different groups of subjects. Note also that the counter-balanced total number of preferences for a particular gap arises from cross-addition of the separate prefererences in separate comparisons.

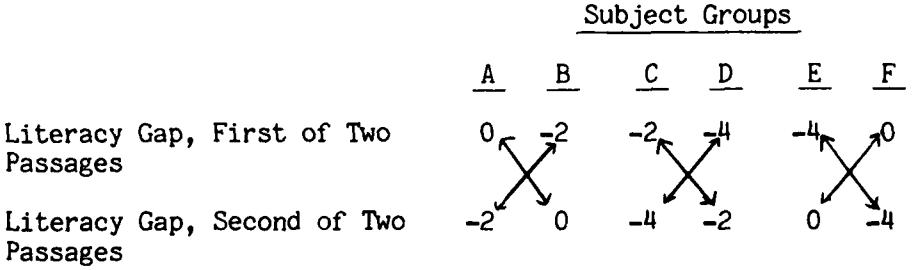


Figure 1. Design for Analysis of Preference Comparisons for Materials Written at Different Literacy Gaps

Procedure

The first step in the testing of subjects involved the administration of the California Achievement Tests (Tiegs & Clarke, 1970): Reading, Level 4 (Grades 6-9), both Vocabulary and Comprehension sub-tests. Subjects followed the standard (published) directions and testing times, completing this phase of experimentation in approximately one hour.

Air Force personnel scored the tests, so that the subjects could be chosen within the desired confidence intervals. Subjects were then assigned to the cells of the 2 x 3 x 3 design according to a randomized scheme prepared beforehand. This meant assigning subjects at the 8th RGL randomly to a literacy gap of 0, -2, and -4, and to a reading time of 30 , 45, or 60 minutes, with the same procedure used for the 10th RGL subjects.

Testing took place in three rooms, one for those given a total of 30 minutes for reading, one for 45 minutes of total reading time, and one for 60 minutes of total reading time. As noted in the Introduction, each of the reading periods were divided into 15-minutes segments, making two for the 30-minute period, three for the 45, and four for the 60-minute period. Tests covering the material read followed each segment. Table 5 presents the times, words read, test items covered, and total comprehension testing time for each segment of each reading period, as well as total time.

Table 5

Data for 15-Minute Reading Time Segments During Comprehension Testing

	Groups*		
	2	3	4
Reading time per segment, minutes	15	15	15
Total reading time, minutes	30	45	60
Number of words read per segment	2,625	1,750	1,313
Total number of words read	5,250	5,250	5,250
Testing time per segment, minutes	20	14	10
Total testing time, minutes	40	42	40
Test items answered per segment	26	17-18	13
Total test items answered	52	52	52
Total experimental time (exclusive of directions), minutes	70	87	100

* Groups labeled in terms of number of 15-minute segments (consequently, there could be no Group 1).

Upon completion of comprehension testing, subjects read two 200-word passages and made judgments on the four preference questions and the fatigue question. Administration of the preference measure was not timed, although subjects completed reading and responding to the questions in approximately 6 minutes. When necessary, subjects were assisted in completing this final phase of the administration of the experimental materials.

Total experimental time, including both comprehension testing and preference measurement, ranged from approximately 1-1/2 hours for the 30-minute group to approximately 2 hours for the 60-minute group.

RESULTS

Reliability Estimates

As noted earlier, comprehension test reliability estmates could not be made very meaningfully on the tryout data for the comprehension tests for the Supervision and Safety and Sanitation contents. Once the comprehension data from the experimental testing became available, however, proper estimates of reliability could be calculated.

Though the comprehension tests were hand-scored, scoring reliability appeared adequate, since a 20% re-check of papers uncovered no errors. Coding and punching of data cards followed scoring, with all statistical analyses performed on Ohio University's IBM S370/158 computer.

Comprehension test reliability estimates were run on the basis of split-half correlations (see Nunnally, 1967, pp. 193-194). For the Supervision content, the correlation turned out to be .82, using the 123 subjects who had read that content. The Safety and Sanitation content yielded a correlation of .88, based on the 117 subjects who had read that content. These figures appear adequate for group testing of the sort done here and compare favorably with those for most reading comprehension tests.

Comprehension Testing

Table 6 presents the comprehension scores of the 240 subjects tested on either the Supervision or the Safety and Sanitation passages.

Table 7 provides descriptive statistics for the Supervision content passage and Table 8 for the Safety and Sanitation content passage, with comprehension scores broken down in terms of:

1. Subject Reading Grade Level (Subject RGL), 8 or 10;

2. Literacy gap, -4 (passage readability grade level four grades higher than Subject RGL), -2 (two grades higher), or 0 (no difference in grades); and

3. Reading time, 30, 45, or 60 minutes.

Table 6

Comprehension Scores for Subjects on the Supervision Passage or the Safety and Sanitation Passage

Supervision Passage		Safety & Sanitation Passage	
Score	f	Score	f
18	1	18	0
19	0	19	0
20	1	20	0
21	0	21	0
22	2	22	1
23	1	23	0
24	2	24	0
25	6	25	0
26	2	26	0
27	7	27	1
28	10	28	0
29	9	29	2
30	10	30	1
31	9	31	3
32	9	32	1
33	7	33	3
34	11	34	3
35	5	35	2
36	5	36	4
37	3	37	4
38	8	38	4
39	5	38	11
40	2	40	13
41	2	41	8
42	2	42	11
43	1	43	5
44	1	44	8
45	0	45	11
46	2	46	6
47	0	47	5
48	0	48	5
49	0	49	2
50	0	50	0
51	0	51	3
	123		117

Table 7

Descriptive Statistics for the Supervision Content
Broken Down by Grade Level, Literacy Gap, and Reading Time

Comprehension Scores

	N	Mean	Standard Deviation
Subject RGL, 8	50	30.24	5.18
Literacy Gap, -4 (12th grade passage)	18	28.94	4.14
Reading Time, 30	6	28.33	4.93
Reading Time, 45	6	27.50	2.59
Reading Time, 60	6	31.00	4.38
Literacy Gap, -2 (10th grade passage)	17	31.59	6.02
Reading Time, 30	5	29.40	6.19
Reading Time, 45	7	32.14	4.22
Reading Time, 60	5	33.00	8.43
Literacy Gap, 0 (8th grade passage)	15	30.27	5.22
Reading Time, 30	4	26.75	6.34
Reading Time, 45	5	29.20	2.39
Reading Time, 60	6	33.50	4.89
Subject RGL, 10	73	33.36	5.15
Literacy Gap, -4 (14th grade passage)	25	32.24	5.72
Reading Time, 30	8	29.88	4.09
Reading Time, 45	8	31.13	4.02
Reading Time, 60	9	35.33	7.21
Literacy Gap, -2 (12th grade passage)	24	33.08	5.18
Reading Time, 30	8	31.13	6.03
Reading Time, 45	8	35.38	4.75
Reading Time, 60	8	32.75	4.33

Table 7 (continued)

	N	Mean	Standard Deviation
Literacy Gap, 0 (10th grade passage)	24	34.79	4.28
Reading Time, 30	8	34.50	4.21
Reading Time, 45	8	36.13	4.12
Reading Time, 60	8	33.75	4.71
Subject RGL, 8	50	30.24	5.18
Subject RGL, 10	73	33.36	5.15
Literacy Gap, -4	43	30.86	5.33
Literacy Gap, -2	41	32.46	5.52
Literacy Gap, 0	39	33.05	5.11
Reading Time, 30	39	30.46	5.42
Reading Time, 45	42	32.31	4.80
Reading Time, 60	42	33.38	5.59
Supervision Content(overall)	123	32.09	5.36

Table 8

Descriptive Statistics for the Safety and Sanitation Content, Broken Down by Grade Level, Literacy Gap, and Reading Time

Comprehension Score

	N	Mean	Standard Deviation
Subject RGL, 8	47	40.19	5.59
Literacy Gap, -4 (12th grade passage)	14	37.79	4.58
Reading Time, 30	6	37.50	3.27
Reading Time, 45	3	39.67	4.73
Reading Time, 60	5	37.00	6.32
Literacy Gap, -2 (10th grade passage)	15	39.53	6.82
Reading Time, 30	4	35.25	10.14
Reading Time, 45	4	37.50	5.92
Reading Time, 60	7	43.14	3.08
Literacy Gap, 0 (8th grade passage)	18	42.61	4.33
Reading Time, 30	5	42.40	3.65
Reading Time, 45	6	44.17	3.71
Reading Time, 60	7	41.43	5.35
Subject RGL, 10	70	41.30	5.14
Literacy Gap, -4 (14th grade passage)	23	41.22	4.43
Reading Time, 30	8	37.75	4.27
Reading Time, 45	8	42.00	2.07
Reading Time, 60	7	44.29	4.23

Table 8 (continued)

	N	Mean	Standard Deviation
Literacy Gap, -2 (12th grade passage)	24	40.71	5.43
Reading Time, 30	8	37.00	4.41
Reading Time, 45	8	41.38	4.24
Reading Time, 60	8	43.75	5.70
Literacy Gap, 0 (10th grade passage)	23	42.00	5.62
Reading Time, 30	7	38.86	6.67
Reading Time, 45	8	42.88	4.39
Reading Time, 60	8	43.88	5.22
Subject RGL, 8	47	40.19	5.59
Subject RGL, 10	70	41.30	5.14
Literacy Gap, -4	37	39.92	4.73
Literacy Gap, -2	39	40.26	5.94
Literacy Gap, 0	41	42.27	5.04
Reading Time, 30	38	38.11	5.42
Reading Time, 45	37	41.73	4.22
Reading Time, 60	42	42.57	5.24
Safety and Sanitation Content (Overall)	117	40.85	5.33

Note the following items in Tables 7 and 8:

1. The Ns for the 8th RGL subjects tend to be low and variable for different cells, compared to the Ns for the 10th RGL subjects.

2. The summary mean values for grade levels, literacy gaps, and reading times fall in the expected directions for both contents, though adjacent differences tend to be small.

3. The overall mean for the Safety and Sanitation content is considerably higher than that for the Supervision content.

The 2 x 3 x 3 analysis of variance on the comprehension scores on the Supervision content yielded the values shown in Table 9.

Table 9

Analysis of Variance of Comprehension Scores on the Supervision Content

Source of Variation	Sum of Squares	df	Mean Square	F	p
Reading Grade Level(RGL)	289.864	1	289.864	11.501	.001*
Literacy Gap (LG)	96.041	2	48.020	1.905	.154
Reading Time (RT)	182.541	2	91.270	3.621	.030*
RGL x LG	47.464	2	23.732	.942	.393
RGK x RT	49.223	2	24.612	.977	.380
LG x RT	103.190	4	25.797	1.024	.399
RGL x LG x RT	93.835	17	23.459	.931	.449
Residual	2646.301	105	25.203		
Total	3510.003	122	28.771		

* Significant beyond .05 level.

Note that the main effects for both reading grade level and for reading time reached significance at the .05 level, but the main effect for literacy gap did not. Note also than none of the interaction effects (two-way or three-way) reached significance at the .05 level.

Analysis of the mean differences for reading grade levels and for reading times indicated that, as anticipated, the subjects at the 10th RGL had significantly higher comprehension scores than the subjects at the 8th RGL. Partition into linear and non-linear components showed that the linear component for reading time was highly significant ($F = 6.16$, $p < .02$). This indicated that as the amount of reading time increases, comprehension scores also increase in a linear fashion.

Table 10 presents the 2 x 3 x 3 analysis of variance of comprehension scores on the Safety and Supervision content Note that the main effect for reading time reached significance at the .05 level, but that the main effects for RGL and literacy gap did not. Note also that none of the interactions reached significance at the .05 level.

Analyses of the mean differences for reading times indicated that as in content 1, the linear component was highly significant ($F = 15.59$, $p < .0001$) with the mean

Table 10

Analysis of Variance of Comprehension Scores on the Safety and Sanitation Content

Source of Variation	Sum of Squares	df	Mean Square	F	p
Reading Grade Level(RGL)	49.288	1	49.288	2.059	.154
Literacy Gap (LG)	113.666	2	56.833	2.375	.098
Reading Time (RT)	423.406	2	211.703	8.845	.001*
RGL x LG	76.896	2	38.448	1.606	.206
RGL x RT	73.959	2	36.980	1.545	.218
LG x RT	103.322	4	25.830	1.079	.371
RGL x LG x RT	84.436	4	21.109	.882	.478
Residual	2,369.420	99	23.934		
Total	3,294.518	116	28.401		

* Significant beyond .05 level.

comprehension score increasing as the amount of reading time increased.

In both contents, there was a nonsignificant trend suggesting that the lower the gap between the subjects' RGL and the readability of the materials, the higher the comprehension score tended to be. This factor was significant for either content, however. Yet the fact that both contents showed the same trend suuggests that this factor may have some effect, albeit a weak one.

Table 11 provides descriptive statistics on the experimental groups by sex and content.

Table 11

Statistics on Experimental Groups' Performance by Sex and Content

	N		Mean		Standard Deviation	
Content	M	F	M	F	M	F
Supervision	101	22	31.81	33.36	5.52	4.48
Safety and Sanitation	77	40	40.94	40.70	5.62	4.79
Totals	178	62				

The analysis of variance presented in Table 12 was conducted because of the observed large differences in content scores and the speculation that sex differences were related to content. As Table 12 shows, however, the main effect for sex as well as the sex by content interaction did not approach significance at the .05 level. The main effect for content, of course, was clearly significant.

Table 12

Analysis of Variance of Experimental Groups by Sex and Content

Source of Variation	Sum of Squares	df	Mean Square	F	p
Sex	0.000	1	0.000	0.027	.870
Content	1.167	1	1.167	69.646	.001*
Sex by Content	.019	1	.019	1.116	.292
Residual	3.954	236	.017		
Totals	5.173	239	.022		

* Significant beyond .05 level.

As mentioned previously, this experiment can additionally be considered as a single 2 x 2 x 3 x 3 design. Combined means for the four experimental variables are presented in Table 13. Because the effect of literacy gap was in the predicted direction but short of significance in each of the individual content analyses, another analysis of variance was performed on the combined data. This was an attempt to increase statistical power. The analysis of variance summary table appears as Table 14. It can be seen that the effect of literacy gap does indeed reach significance in this analysis, and that content does not interact with any other experimental variable.

The effect of reading time is significant for the combined data, as it was in each of the individual analyses. While increased reading time led to increased comprehension scores, efficiency dropped off as reading time increased. In this study, efficiency may be roughly evaluated by simply dividing the comprehension scores by the amount of study time for the several groups. Table 15 provides a comparison of these figures for the two contents used. Note that the number of items answered correctly per minute of study time drops off rather rapidly as the time increases from 30 minutes to 45 minutes to 60 minutes. Of course, the nature of the test used sets an upper limit on the number possible and this cannot therefore be taken in quite so straightforward a manner as the figures would suggest.

Table 13

Mean Comprehension Scores Broken Down By Content, Subject RGL, Literacy Gap and Reading Time

Combined Means

Source of Variation	N	Mean	%
Overall	240	36.36	69.9%
Content			
Supervision	123	32.09	61.7%
Safety	117	40.85	78.6%
Subject RGL			
8th	97	35.06	67.4%
10th	143	37.25	71.6%
Lit Gap			
-4	80	35.05	67.4%
-2	80	36.26	69.7%
0	80	37.78	72.7%
Reading Time			
30	77	34.24	65.8%
45	79	36.72	70.6%
60	84	38.	73.1%

Table 14

Analysis of Variance of Comprehension Scores on Combined Contents

Source of Variation	Sum of Squares	df	Mean Square	F
Main Effects	5660.324	6	943.387	38.371**
Content	4510.793	1	4510.793	183.471**
RGL	291.690	1	291.690	11.864**
Gap	188.149	2	94.075	3.826*
Reading Time	572.906	2	286.453	11.651**
Two-Way Interactions	239.359	13	18.412	0.749
Content X RGL	37.292	1	37.292	1.517
Content X Gap	31.926	2	15.963	0.649
Content X Time	28.082	2	14.041	0.571
RGL X Gap	31.140	2	15.570	0.633
RGL X Time	25.193	2	12.597	0.512
Gap X Time	77.530	4	19.383	0.788
Three-Way Interactions	418.023	12	34.835	1.417
Content X RGL X Gap	104.131	2	52.066	2.118
Content X RGL X Time	105.783	2	52.892	2.151
Content X Gap X Time	132.597	4	33.149	1.348
RGL X Gap X Time	100.265	4	25.066	1.020
Four-Way Interactions	77.996	4	19.499	0.793
RGL X Content X Gap X Time	77.996	4	19.499	0.793
Explained	6395.703	35	182.734	7.433
Residual	5015.508	204	24.586	
Total	11411.211	239	47.746	

* $p < .05$
** $p < .001$

Table 15

Average Number of Items Answered Correctly Per Minute for the Several Study Times Used by the Experimental Subjects

Study Times	Supervision Content	Study and Sanitation Content
30 Minutes	1.02 items/min	1.27 items/min
45 Minutes	.72 items/min	.93 items/min
60 Minutes	.56 items/min	.71 items/min

Preference Measurement

As noted earlier, the preference questions asked subjects to indicate whether they preferred the first or second of two passages of approximately 200 words each. In half of the paired passages, the first passage presented was easier, i.e., had a lower literacy gap, and in the other half the first passage was harder, i.e., had a higher literacy gap. All possible pairs of literacy gaps appeared, given to approximately equal numbers of subjects. Figure 2 presents the composition and description of the pairs of passages used, i.e., first passage easier or second passage easier. These descriptions appear in Table 16, which presents the number of subjects who selected the first or second passage for each preference question, or who indicated no difference between them. The questions themselves asked subjects to decide which one of the pairs of passages seemed: (a) easier, (b) more informative, (c)more interesting and (d) clearer.

	First Passage	Second Passage	Description of Pairs
Literacy Gap	0	-2	First passage easier
Literacy Gap	0	-4	First passage easier
Literacy Gap	-2	-4	First passage easier
Literacy Gap	-4	-2	Second passage easier
Literacy Gap	-4	0	Second passage easier
Literacy Gap	-2	0	Second passage easier

Figure 2. Composition and Description of Pairs of Passages Used

Table 16

Numbers of Subjects Who Selected the First or the Second Passage or Who Indicated No Difference Between Them

Description Of Pair	Supervision Content (N = 117)			Safety and Sanitation Content (N = 123)		
	Subjects Judging Easier			Subjects Judging Easier		
	First Pass.	Second Pass.	Equal	First Pass.	Second Pass.	Equal
First Easier	12	7	73	8	6	95
Second Easier	10	15	(62%)	2	12	(77%)
	Subjects Judging More Informative			Subjects Judging More Informative		
First Easier	18	19	38	16	32	28
Second Easier	20	22	(32%)	14	33	(23%)
	Subjects Judging More Interesting			Subject Judging More Interesting		
First Easier	14	24	35	9	37	26
Second Easier	16	28	(30%)	14	37	(21%)
	Subjects Judging Clearer			Subjects Judging Clearer		
First Easier	16	13	60	13	12	68
Second Easier	8	20	(51%)	9	21	(55%)

Note the following in Table 16:

1. Subjects who had read the Supervision content for purposes of comprehension testing, instead read passages from the Safety and Sanitation content for purposes of preference judgments, and vice versa. Consequently the N's for the two contents for comprehension testing, i.e., 123 and 117 respectively, are reversed for preference measrement.

2. In 338 out of 537 judgments (63%), subjects selected the second passage, regardless of the readability of the passages or the questions asked.

3. In 423 out of 960 judgments (44%), subjects failed to see a difference between the two passages in a pair.

Fortunately, the counterbalanced design used for preference measurement made it possible to eliminate the order effect noted in item 2 above. Similarly, the design helps, at least, to get around the problem of large numbers of subjects seeing no difference between the two passages of a pair. The analysis makes use of cross-addition so that preference for an easier passage, when it appeared first in a pair can be combined with the preference for an easier passage, when it appeared second. Figure 1 portrays this procedure and Table 17 provides the actual comparisons for the data in Table 16. Percentages are given for the numbers selecting the easier pasages as opposed to the harder of the pairs.

Note the following in Table 17:

1. Considering only those subjects who perceived differences in the passages, 61% for one content and 71% for the other content correctly judged the easier passages to be easier.

2. Considering only those subjects who perceived differences in the passages, 63% for one content and 62% for the other content correctly judged the easier passages to be clearer.

3. Considering only those who perceived differences in the passages, subjects judged the pairs of passages about equally informative, favoring the easier by percentages by only 51% and 52% for the two contents. These percentages support the equivalence of the "information content" as opposed to the readability or style difficulty of the several versions, as judged during the preparation of the versions.

4. Considering only those who perceived differences in the passages, subjects judged the pairs of passages about equally interesting, favoring the easier by 51% for one content and the harder by 53% (the inverse of 47%) for one content. These percentages again support the equivalence of the "information content" as opposed to the readability of the several versions.

Table 17

Cross-Addition of Preference Judgments and Percentages Selecting the Easier Passages as Opposed to the Harder of the Pairs

Descrip of Pair	Supervision Content: Subjects Judging Easier — First Passage	Second Passage	Easier Percentage	Safety and Sanitation Content: Subjects Judging Easier — First Passage	Second Passage	Easier Percentage
First Easier Second Easier	12	7	27/44 = 61%	8	6	20/28 = 71%
	10	15		2	12	
	Subjects Judging More Informative			Subjects Judging More Informative		
First Easier Second Easier	20	19	42/81 = 52%	16	32	49/95 = 52%
	20	22		14	33	
	Subjects Judging More Interesting			Subjects Judging More Interesting		
First Easier Second Easier	14	24	42/82 = 51%	9	37	46/97 = 47%
	16	28		14	37	
	Subjects Judging Clearer			Subjects Judging Clearer		
First Easier Second Easier	16	13	36/57 = 63%	13	12	34/55 = 62%
	8	20		9	21	

In view of the small numbers of judgments involved for each content separately, and the similarity of the preferences for the two contents, the figures for the two contents have been combined for purposes of significance testing. Table 18 presents the results of these tests, combining the two contents for each of the questions.

Table 18

Chi-Square Tests* of Preference Judgments on Combined Contents

Description of Passage	Subjects Judging Easier: First Passage	Second Passage	Totals	
First Passage Easier	20	13	33	
Second Passage Easier	12	27	39	$\chi^2 = 6.42$,
Totals	32	40	72	$p < .02$

	Subjects Judging More Informative: First Passage	Second Passage	Totals	
First Passage Easier	36	51	87	
Second Passage Easier	34	55	89	$\chi^2 = .19$,
Totals	70	106	176	n.s.

	Subjects Judging More Interesting: First Passage	Second Passage	Totals	
First Passage Easier	23	61	84	
Second Passage Easier	30	65	95	$\chi^2 = .38$,
Totals	53	126	179	n.s.

	Subjects Judging Clearer: First Passage	Second Passage	Totals	
First Passage Easier	29	25	54	
Second Passage Easier	17	41	58	$\chi^2 = 6.82$,
Totals	46	66	112	$p < .01$

* Note that the chi-square test is equivalent to testing for a single proportion when there are only two categories; see Hays, 1963, page 585.

The tests given in Table 18 lend statistical support to the comments above. That is, the easier of a pair of passages was judged significantly easier and clearer by subjects who perceived some difference. On the other hand, the easier of a pair of passages was judged neither more informative nor more interesting.

The final analysis involves the answers to the fifth and last of the "preference" questions, concerning subject fatigue. Table 19 presents these data. Note that N = 244,

indicating that the four of the five subjects removed from the data for the purposes of comprehension testing and preference measurement have been included here for increased N. As Table 19 indicates, few of the subjects indicated they were tired at the end of the experimental session. Analysis of possible decrement in comprehension score beyond the first reading-test period corroborate these data, since no clear-cut fall-off in scores appeared. Instead, the analyses yielded much the same conclusions as the 2 X 3 X 3 analyses of the total scores. Consequently, these analyses have not been presented here.

Table 19

Answers to the Question Concerning Subject Fatigue (N = 244)

Judgements	N	%
Not at all tired	100	41
Beginning to feel tired	98	40
Pretty tired	38	16
Very tired	8	3
Total	244	

DISCUSSION

Introductory Remarks

Many studies have been done on the effects of modified readability upon student comprehension. What can this study offer in the way of added knowledge? Certain desirable characteristics make it unique, and its problems as well as its implications should therefore be of interest. A summary of these characteristics follows:

1. The study involved operational Air Force career development materials rather than materials specially created for the purpose of experimentation, as is often the case. Consequently, the materials should have a certain face validity. Furthermore, the results should generalize to other such materials to an extent not possible otherwise.

2. Two contents, or topics were examined in the study rather than only one. The complexity of human differences in background interests, attitudes and capabilities interacting with the great variety of written materials made obvious the need to use more than one topic. This study as well as that of Denbow (1973) clearly support this need in showing that the same readability treatment may have different effects with two different contents or topics.

3. Readability was varied over long passages. Readers typically face long bodies of text, yet most experimental workers, for reasons of time, effort, and cost, limit themselves to short passages. In some cases, they use single sentences, raising serious questions about the ability to generalize from the results.

4. The readability versions were constructed with great care, in order both to specify clearly for others how to make such changes in readability and to make possible the clear interpretation of any cause-effect relationships which might be found. Note especially the following:

a. This study used adult subjects with limited reading skills who might be assumed to encounter problems in dealing with typical Air Force reading materials.

b. The subjects took a reading test to determine their reading skill levels and were selected to fall within specified confidence intervals within the specified RGLs.

c. Materials at several grade levels at or beyond the tested skill levels of the subjects were developed and used to create specific "literacy gaps." These gaps were those most likely to be encountered by Air Force personnel with limited reading skills.

d. Reading times were varied, to study the effect of added reading time upon comprehension. Three levels of reading time were used.

e. Changes in readability were made according to clearly specified word and sentence suggestions, as contained in A Manual for Readable Writing (Klare, 1975).

f. Readability levels were carefully determined by using the Kincaid version of the Flesch Reading Ease formula (Kincaid, Fishburne, Rogers, and Chissom, 1975). Individual sections, as well as the passages as a whole, were controlled for readability.

g. Important controls were applied to the several readability versions to increase the precision of the experiment. These included: length of the versions, information content (as opposed to style difficulty or readability), and retention of technical terms.

5. The comprehension tests were constructed carefully from a large item pool in order to achieve adequate reliability and sensitivity of measurement. Note especially that the following was accomplished by careful selection from a large body of trial items.

a. The items were spread across the content of the 5,250-word passages, and were keyed to the content. Thus they could be divided into sub-sections which corresponded to the sub-sections of text which was read during the three experimental reading times.

b. Item analysis procedures were employed to determine difficulty levels and item-test correlations. This resulted in a comprehension test with high reliability.

c. Versions of the comprehension test were prepared so that, with "information content" constant, the versions corresponded in readability to the readability of the experimental passages. The common complaint about multiple-choice tests that one "can write easy items about difficult content and difficult items about easy content" was thus addressed.

6. Preference measures were included in the study so that reader feelings about the readability versions could be assessed. These measures provide a check on:

a. The judged ease and clarity of the several versions; and

b. The judged "information content" and interest value of the several versions.

7. The objectivity of the study was maximized by completion of the work in three separate locations. The design of the study and the analysis of data were performed primarily by personnel from Ohio University. The writing of the experimental versions and the overall supervision on the project were carried out primarily by personnel from Defense and Electronic Systems Center, Integrated Logistics Support Division, of Westinghouse Electric Corporation, in Hunt Valley, Maryland. The development of the comprehension test and the experimental testing was handled by personnel from Measurement Research Center, Westinghouse Electric Corporation, in Iowa City, Iowa. Personnel from the Air Force Human Resources Laboratory Technical Training Division at Lowry AFB monitored the project and assisted in many phases of its execution. These cooperative yet independent efforts provided some insulation against the frequent charge of experimenter bias in the direction of "finding what one wants to find." With this brief introduction as preface, the results of the study can be discussed and their significance assessed.

Comprehension Testing

The summary means for both passages, or contents, fell in the expected direction; specifically (a) lower means were found for 8th RGL subjects than for 10th RGL subjects; (b) lower means were found for a literacy gap of -4 than of -2, and for a literacy gap of -2 than of 0; and (c) lower means were found for a reading time of 30 than one of 45 minutes than one of 60 minutes. Yet, for the Supervision passage, differences significant at or beyond the .05 level emerged for only reading grade level and reading time. For the Safety and Sanitation content the only difference significant at the .05 level or beyond turned out to be that for reading time.

While all the factors were significant for the combined data, the size of the effects were small. Why should this be so? The answer must be somewhat speculative at this point, but several hypotheses seem relevant. Of course, the answer might well lie in some combination of these reasons:

1. As noted earlier in this report, the number of subjects available at the desired RGLs turned out to be smaller than desired. Special efforts were made by Air Force personnel to obtain additional subjects, but without complete sucess. For example, almost no 6th RGL subjects could be located. Though not without some possible satisfaction for Air Force personnel (who appear to be getting recruits with high level reading skills), this event forced a revision of the original experimental design and reduced the power of the statistical tests, particularly the case of the RGL variable. This may well have helped to

account for the lack of significance for this variable in the Safety and Sanitation content. Note, in this connection, the mean score for several subjects at the 6th RGL who took different levels of the Safety and Sanitation test was 30.67. This compares with mean scores over all test levels of 40.19 and 41.30 respectively.

In addition to the above problem, the relatively small and variable numbers of subjects at the 8th RGL in the various cells of the design contributed additional problems. This certainly might have played a part in the lack of significance for the RGL variable. Recall, in connection with this, that the mean for these subjects came out higher than that given in the norms tables, i.e., 53.2 versus 52. On the other hand, the mean for the subjects at the 10th RGL fell in the opposite direction from the mean in the norms tables, i.e., 62.3 versus 63.5. This restriction of the difference beween the mean grade levels of subjects, though probably not serious, may at least have contributed to the lack of significance for the Safety and Sanitation content.

And, of course, the practical needs and considerations of training ruled out obtaining the ideal of 10 subjects per cell at even the 10th RGL. This should not be taken as criticism of the efforts made to obtain subject, because these efforts could not be faulted; rather, it should be taken as one possible contributor to the results observed.

A related matter concerns the inability to stay within the desired 95% confidence interval in selecting subjects at the 8th and 10th RGL's. The interval had to be expanded to 99.99999% in the former case and to 99.9% in the latter case. The use of confidence intervals remains an advance over many studies using subject variables such as reading grade level, since such a step is seldom taken. Nevertheless, use of the broader intervals meant at least somewhat greater error variance and may thus have contributed somewhat to the lack of significance of the several variables already noted.

2. The experimental materials and the comprehension tests must always be considered possible contributors when non-significant results are found. Examination, however, revealed no obvious flaws in these areas.

For one thing, the desired readability levels of the experimental versions were carefully adhered to, not only for the passages as a whole but also for the separate segments of the passages. For another, the changes made were not simply "index" changes, but rather "causal" changes, based upon the psycholinguistic findings summarized in A Manual for Readable Writing. (See Klare, 1976, for a discussion of the index causal variable issue.) Furthermore,

careful controls were applied for length and information content.

Similarly, the comprehension tests were carefully constructed from the tryout data on a large number of trial items. The sub-tests were arranged to correspond to the sub-sections read during the three different reading times. And the special step of matching the readability of the test versions to that of the text they covered removed another possible criticism of multiple-choice tests.

3. As noted in the Introduction, reader motivation may have played a part in reducing the likehood of significant results here, especially as regards the literacy gap variable. In the paper mentioned earlier (Klare, 1976), comparisons were made of where modified readability produced versus where it failed to produce significant differences in comprehension scores. This examination revealed that where a raised level of motivation interacting with testing has an effect, the chances for significant differences in comprehension were reduced. These conditions prevailed in this experiment. For one thing, the test situation itself tended to raise motivation somewhat, so repeated testing, as done here, made such a rise more likely. For another, the liberal reading times, especially for the 45- and the 60-minute reading groups of readers, provided an opportunity for this motivation to be effective. As noted earlier, mean comprehension scores did increase as the amount of reading time increased.

Experimental studies support this motivation interpretation. As noted, Fass and Schumacher (in press) have shown that increased reward can significantly reduce the effect of readablity upon comprehension scores. McLaughlin (1966) has shown the same thing for threat. And Denbow (1973) has demonstrated that even the motivation as measured by an expressed preference for content can have this kind of effect.

Preference Measurement

The preference questions produced summary score values in the predicted directions for both contents: Supervision and Safety and Sanitation. The more readable of the pairs of passages were judged both easier to read and clearer. This finding supports the notion that, had other conditions (especially the increased level of motivations) not tended to reduce the likelihood of significance, readability might well have been more clearly effective in terms of comprehension scores.

On the other hand, the summary scores also showed that the various readability versions were virtually equal in

terms of information content and interest-value. This finding supports the notion that the controls on content (i.e., through use of judges) proved effective and that the versions differed only in readability. However, the large number of judgements that the pairs of passages did not differ proved disappointing. As to why this could have happened, several possibilities can be suggested:

1. In retrospect, a forced-choice arrangement (i.e., not allowing a no-difference option) may well have been a better approach despite the error variance it may have added. Only further research could answer such a question.

2. Another possibility must be that the pairs of preference passages were quite short--only about 400 words. In a similar study of the judgments of Air Force personnel (Klare, Mabry, Gustafson, 1955), the passages used were three times as long, and the results were more clear-cut. Perhaps judgments of ease and clarity of reading seem difficult enough to readers that more text would be helpful. Short passages were used here only because added experimental time was undesirable.

3. A related possibility concerns the time when the preference measurement took place--at the end of the experimental session. This could conceivably have dulled the subjects' ability to make such judgments. On the other hand, few subjects reported fatigue, and the smaller number of no-difference judgments for the information content and interest-value questions does not support this explanation.

In sum, the length-of-passage explanation seems most likely. At any rate, the counterbalanced design meant that the hypothesis could be tested in spite of the large number of no-difference judgments, and could yield useful explanatory data.

CONCLUSIONS

1. Literacy gap produced a small but significant effect upon comprehension scores under the conditions of this study, using relatively long passages of approximately 5000 words. One possibility suggested by previous readability research is that the repeated testing during the experiment induced a high level of motivation in the subjects, and that the liberal reading and testing times allowed this motivation to reduce the effect of readability upon comprehension scores. Perhaps, too, the scarcity of appropriate subjects at the lower reading levels contributed to the attenuation of the effect.

2. Increasing reading time, for the range of times used here, appears to increase the text comprehension scores of readers. However, the relation between reading time and comprehension scores is such that subjects given more time learn less efficiently (i.e., learn less per unit time). The effect of added reading time does not appear to vary with level of literacy gap.

3. A majority of subjects did not perceive differences between pairs of short passages of approximately 200 words written at different levels of readability. Those subjects who did indicate a preference, however, significantly favored the more readable of the pair. Previous readability research suggests that this effect may have been more marked if longer passages had been used.

IMPLICATIONS FOR FURTHER STUDY

These recommendations lead to some implications for further research, as suggested below.

1. Measures of the efficiency of learning from prose have recently been given renewed emphasis. Arkes, Schumacher, and Gardner (1976) used this type of measure, and Faw and Waller (1976) have re-evaluated a number of studies by means of such a measure. They have found that many experiments which purport to show increased learning have actually showed little if any increase in amount learned per unit of study time. In other words, the experimental conditions simply required more time of the subjects for the amount learned. Such studies raise some questions for future experimentation.

a. For one thing, number of readings might be considered as a possible variable in some future work. Efficiency could be examined under such conditions also.

b. What is a "desirable" level of comprehension? If a high level is desired, perhaps a long study time is justified if the proper motivational conditions can be developed and mantained so that the study time is really effectively used. And a major question, of course, concerns whether such conditions would remain effective over extended periods.

c. Perhaps a more practical question concerns the amount likely to be learned under conditions of "typical" motivation and "typical" study time. This matter will be considered more fully below.

2. Concern for the question of whether one can generalize from the results of experimental studies has been around for a long time. Relatively satisfactory answers are available for the question of generalizing from samples of subjects to a population. Some attention has also been given to the matter of generalizing to a language population, but there still is disagreement among statisticians concerning the best way to handle this problem. For example, see Coleman (1964), Clark (1973), and the series of responses engendered by Clark's article; see Wike & Church (1976) and Clark (1976). Perhaps the least attention has been given to the problem of generalizing from the results found under experimental conditions to the real world. A notable recent exception to this has been the article of Gadlin and Ingle (1975). And, of course, Webb, Campbell, Schwartz, and Sechrest (1966) have pioneered in the answer to such concerns in their book on unobtrusive measures. In the field of readability, Klare (1976) has raised the same concern. He has pointed out that a more

nearly ideal answer to the effects of readability must come from field studies using unobtrusive measures. Such studies can never be easy to do, but when they have been done, the results have been more clear-cut than for experimental studies. This appears to be due, in large measure, to the motivational variable. Where practical concerns predominate, the ideal conditions for testing must be typical levels of motivation and typical conditions of study. Finding such conditions and creating a field study with truly unobtrusive measures, though difficult, would not be impossible. If, as research has shown, preferences constitute a major determiner of what and how much one will read, comprehend and retain, the present results support the need for such future work. This kind of study of the effect of readability upon comprehension would thus appear to be one of the logical next steps for Air Force research in this area.

RECOMMENDATIONS FOR THE AIR FORCE

Recommendation 1. The relationship of literacy gap specifically to job performance should be examined before major efforts are undertaken to rewrite Air Force materials for greater ease of reading.

Recommendation 2. Efforts to improve readability of materials might best be directed at populations and situations where motivation and interest are unlikely to be high.

Recommendation 3. Increasing reading time would seem to be a reliable and straight-forward way to increase text comprehension under conditions of high motivation. However, because of the decreased learning efficiency that this method is likely to induce, a careful analysis of whether the gain in comprehension is worth the extra expenditure of reading time should first be performed. It is clear that there is some point in any interaction of reader and text where no amount of further reading time improves comprehension.

REFERENCES

Arkes, H. R., Schumacher, G. M., & Gardner, E. T. Effects of Orienting Tasks on the Retention of Prose Materials. Journal of Educational Psychology, 1976, 68, 536-545.

Burkett, J. R. A Review of Research and Development on Literacy and Technical Writing in the Air Force. Paper presented at the Conference on Reading and Readability Research in the Armed Services, U. S. Naval Post-Graduate School, Monterey, CA, October 28-30, 1975.

Clark, H. H. The Language-as-Fixed-Effect Fallacy: A Critique of Language Statistics in Psychological Research. Journal of Verbal Learning and Verbal Behavior, 1973, 12, 335-359.

Clark, H., Cohen, J., Smith, J. E. K., & Kepel, G. Discussion of Wike and Church's Comments. Journal of Verbal Learning and Verbal Behavior, 1976, 15, 257-266.

Coleman, E. B. Generalizing to a Language Population. Psychological Reports, 1964, 14, 219-226.

DeGuelle, J. R. USAFI Achievement Test III Diagnosis of Reading Deficiencies for Air Force Personnel. Unpublished manuscript, Aerospace Defense Command, January 1975.

Denbow, C. J. An experimental Study of the Effect of a Repetition Factor on the Relationship Between Readability and Listenability. Unpublished doctoral dissertation, Ohio University, 1973.

Entin, E. B., & Klare, G. R. Some Relationships Between Readability, Cloze, and Multiple-Choice Scores on Published Reading Comprehension Tests. Journal of Reading Behavior, in press.

Fass, W., & Schumacher, G. M. The Effects of Motivation, Subject Activity, and Readability on the Retention of Prose Materials. Journal of Educational Psychology, in press.

Faw, H. W., & Waller, T. G. Mathemagenic Behaviors and Efficiency in Learning from Prose. Review of Educational Research. 1976, 46, 691-720.

Frase, L. T., & Schwartz, B. J. Typographical Cues that Facilitate Comprehension. Murray Hill, NJ: Bell Laboratories (undated).

Gadlin, H., & Ingle, G. Through the One-Way Mirror: The Limits of Experimental Self-Reflection. American Psychologist, 1975, 30, 1003-1009.

Hays, W. L. Statistics for Psychologists. New York: Holt, Rinehart and Winston, 1963.

Kincaid, J. P., Fishburne, R. P., Rogers, R. L., & Chissom, B. S. Derivation of New Readability Formulas (Automated Readability Index, Fog Count, and Flesch Reading Ease Formula) for Navy Enlisted Personnel. Research Branch Report 8-75. Chief of Naval Technical Training, Naval Air Station, Memphis (75), Millington, TN 38054. 1975.

Klare, G. R. The Measurement of Readability. Ames, Iowa: Iowa State University Press, 1963.

Klare, G. R. Words, Sentences, and Readability. Technical Report No. B.1.1, Project CREATES. Cambridge, MS: Harvard Computer-Aided Instruction Laboratory, 1969. Mimeo.

Klare, G. R. Assessing Readability. Reading Research Quarterly, 1974-1975, 10. 62-102.

Klare, G. R. A Manual for Readable Writing. Glen Burnie, Maryland: REM Co., 119A Roesler Road, 1975.

Klare, G. R. A Second Look at the Validity of Readability Formulas. Journal of Reading Behavior, 1976, 8, 129-152.

Klare, G. R., Mabry, J. E., & Gustafson, L. M. The Relationship of Style Difficulty to Immediate Retention and to Acceptability of Technical Material. Journal of Educational Psychology, 1955, 46, 287-295.

Klare, G. R., Sinaiko, H. W., & Stolurow, L. M. The Cloze Procedure: A Convenient Readability Test for Training Materials and Translations. International Review of Applied Psychology, 1972, 21, 77-106.

Madden, H. L., and Tupes, E. C. Estimating Reading Level from the AQE General Aptitude Index. PRL-TR-66-1, AD-632182. Lackland Air Force Base, Texas, Personnel Research Laboratory, February 1966.

McLaughlin, G. H. Comparing Styles of Presenting Technical Information. Ergonomics, 1966, 9, 257-259.

Miller, L. R. A Comparative Analysis of the Predictive Validities of Four Readability Formulas. Unpublished doctoral dissertation, Ohio University, 1972.

Mockovak, W. P. An investigation and methodologies for determining the reading skills and requirements of Air Force career ladders. AFHRL-TR-73-53, AD-777 834. Lowry Air Force Base, CO, Technical Training Division, Air Force Human Resources Laboratory, January 1974. (a)

Mockovak, W. P. Literacy Skills and Requirements of Air Force Career Ladders. AFHRL-TR-74-90, AD-A008 770. Lowry Air Force Base, CO, Technical Training Division, Air Force Human Resources Laboratory, December 1974.

Nunally, J. C. Psychometric Theory. New York: McGraw-Hill, 1967.

Sticht, T. G. (Ed) Reading for Working. Alexandria, VA: Human Achievement Tests. Monterey, CA: CTB McGraw-Hill, Del Monte Research Park, 1970.

Tiegs, E. W., & Clarke, W. W. Examiner's manual, California Achievement Tests. Monterey, CA: CTB McGraw-Hill, Del Monte Research Park, 1970.

Webb, E. J., Campbell, D. T., Schwartz, R. D., & Sechrest, L. Unobtrusive Measures: Nonreactive Research in the Social Sciences. Chicago: Rand McNally, 1966.

Wike, E. L., & Church, J. D. Comments on Clark's "The Language-as-Fixed-Effect Fallacy." Journal of Verbal Learning and Verbal Behavior, 1976, 15, 249-256.

APPENDIX A - SAMPLES FROM THE SUPERVISION PASSAGES

1. Sixth Grade Version

Basic Needs of People. The fact that a worker may reach a sufficient production level doesn't mean he will stay at that level without more attention. As you well know, many things cause a worker to "let down" in his work. For seemingly no reason a worker may change quickly from a satisfactory, content performer into an unhappy, low producer. One of the duties of a supervisor is to learn what these reasons are and help that worker on the way back to high output. This is a tough job since there are no hard and fast rules that work for all. Workers aren't machines. You can't just push a button and make them do what you want them to do. You can't just look at them and find out why they don't run right. Workers are people with ambitions that can be stirred and pride that can be hurt. They have nerves that can be shattered and hopes that can come true. This makes workers complex and hard to understand; but, you must understand them if you are to build a content, helpful and productive work force. To help you grasp what makes people act as they do, you need to know the basic needs of people. These needs are: recognition, opportunity, security, and a sense of belonging.

2. Eighth Grade Version.

Basic Needs of People. The fact that a worker may reach a satisfactory production level doesn't mean he will remain at that level without further attention. As you well know, many things cause a worker to "let down" in his performance. For seemingly no reason, a worker may change overnight from a satisfactory, satisfied performer into an unhappy, low producer. One of the duties of a supervisor is to find out what these reasons are and to help that worker on the way back to high performance. This is a difficult job since there are no hard and fast rules that work for all. Workers aren't machines so you can't just push a button and make them do what you want them to do. You can't just look at them and determine why they don't run right. Workers are people with ambitions that can be stirred, pride that can be hurt, nerves that can be shattered, and hopes that can come true. This makes workers complex and hard to understand. But you must understand them if you are to build a content, cooperative, and productive work force. To help you understand what makes people act as they do, you need to know the basic needs of people, These needs are recognition, opportunity, security, and a feeling of belonging.

3. Tenth Grade Version.

Basic Needs of People. The fact that a worker may reach a satisfactory production level doesn't mean he is going to remain at that level without further attention. As you know, many factors cause a worker to "let down" in his performance. For no apparent reason, a worker may change overnight from a satisfactory, satisfied performer into an unhappy, low producer. One of the obligations of a supervisor is to find out what these reasons are and to assist that worker on the way back to high performance. This is a difficult job because there are no hard and fast rules that work for everyone. Workers aren't machines, so you can't merely push a button and make them do what you want them to do. You can't just look at them and determine why they don't run properly. Workers are people with ambitions that can be stirred, pride that can be hurt, nerves that can be shattered, and hopes that can be realized. This makes workers complex and difficult to understand; however, you must understand them if you are to develop a satisfied, cooperative, and productive work force. To help you understand what makes people act as they do, it is necessary for you to know the basic needs of people; recognition, opportunity, security, and a feeling of belonging.

4. Twelfth Grade Version

Basic Needs of People. The fact that a worker may attain a satisfactory production level doesn't indicate that he is going to remain at that level without additional attention. As you well realize, numerous factors cause a worker to "let down" in his performance. For no apparent reason, a worker may convert overnight from a satisfactory, satisfied performer into an unhappy, low producer. One of the responsibilities of a supervisor is to uncover what these reasons are and to assist that worker onto the pathway back to high performance. This is a difficult job because there are no hard and fast rules that work for everyone. Workers aren't machines so you can't simply push a button and make them do what you want them to. You can't just look at them and determine why they don't run properly. Workers are people with ambitions that can be invigorated, pride that can be injured, nerves that can be shattered, and hopes that can be realized. This makes workers complex and difficult to understand, but you must understand them if you are to develop a satisfied, cooperative, and productive work force. To help you understand what makes people act as they do, it is necessary for you to know the fundamental needs of people: recognition, opportunity, security, and a feeling of belonging.

5. Fourteenth Grade Version

Basic Needs of People. The fact that a worker may attain a satisfactory production level doesn't indicate that he is going to remain at that level without additional attention. As you well realize, numerous factors cause a worker to "let down" in his performance. For apparently no reason, a worker may transform overnight from a satisfactory, satisfied performer into an unhappy, low producer. One of the responsibilities of a supervisor is to discover what these reasons are and to assist that worker onto the pathway back to superior performance. This is a difficult undertaking because there are no hard and fast rules that work for everyone, and workers aren't machines that can be manipulated into doing whatever you say by merely pushing a button. You can't simply observe them and determine why they don't function properly. Workers are individuals with ambitions that can be invigorated, pride that can be injured, nerves that can be shattered, and hopes that can be realized. This makes workers complex and difficult to understand, but you must understand them if you are to develop a satisfied, cooperative, and productive work force. To assist you in understanding what makes people act as they do, it is necessary for you to know the fundamental needs of people: recognition, opportunity, security, and a feeling of belonging.

APPENDIX B - SAMPLES FROM THE SAFETY AND SANITATION PASSAGE

1. Sixth Grade Version

Here are some interesting and important facts about rats that should be kept in mind when rodent-proofing buildings. (1) Rats can enter holes as small as 1/2 inch wide, (2) rats can climb better straight up and down. (3) Rats can climb pipes 4 inches around or smaller. (4) Rats can jump 3 feet high from a flat surface. (5) Rats can jump 4 feet across a flat surface. (6) Rats can jump 8 feet from an elevated position. (7) Rats can fall 50 feet without hurting themselves. Also, rats prefer to travel and hunt for food at night. They are creatures of habit and almost always travel from their nest to their food sources and to the outside over the same paths. Maybe for protection, their paths usually are in narrow, out-of-the-way places, like overhead pipes and beams, or along walls. When rats run from place to place, they hug the wall. Rat runs are easy to find because dirt and oil from their hair rub off and blacken the surfaces they touch.

2. Eighth Grade Version

Some interesting and important facts about rats which should be kept in mind when rodent-proofing buildings are: (1) rats can enter holes as small as 1/2 inch in diameter; (2) rats can climb better vertically; (3) rats can climb pipes 4 inches in diameter or smaller; (4) rats can jump 3 feet high from a flat surface and they can jump 4 feet horizon- tally; (5) rats can jump 8 feet from an elevated position; and (6) rats can fall 50 feet without injuring themselves. Also, rats prefer to travel and hunt for food at night. They are creatures of habit and almost always travel from their nest to their food sources and to the outside over the same paths. Perhaps for protection, their paths usually are in narrow, out-of-the-way places, such as overhead pipes and beams, or along walls. When rats run from place to place, they hug the wall. Rat runs are easy to find because dirt and oil from the hair on the rats rub off and blacken the surfaces that they touch.

3. Tenth Grade Version

Some interesting and significant facts about rats which you should be aware of when rodent-proofing buildings are: (1) rats can enter into holes as small as 1/2 inch in diameter; (2) rats can climb better vertically and they can climb pipes 4 inches in diameter or smaller; (3) rats can jump 3 feet high from flat surfaces, and they can jump 4 feet horizontally; and (4) rats can jump 8 feet from an elevated position or they can fall 50 feet without injury to themselves. Also, rats prefer travelling and searching for food at night. They are creatures of habit and almost always travel from their shelter to their food sources and to the outside over the same pathways. Perhaps for protection, their pathways usually are in narrow, out-of-the-way places, such as overhead pipes and beams, or along walls. When rats are running from one location to another, they hug the wall. Rat runs are easy to find because dirt and oil from the hair on the rats rub off and blacken the surfaces that they touch.

4. Twelfth Grade Version

Some interesting and significant facts concerning rats which should be remembered when rodent-proofing buildings are: (1) rats can enter holes as small as 1/2 inch in diameter; (2) rats can climb better vertically, and they can climb pipes 4 inches in diameter or smaller; (3) rats can jump 3 feet high from flat surfaces, and they can jump 4 feet horizontally; (4) rats can jump 8 feet from an elevated position, and they can fall 50 feet without injuring themselves. Additionally, rats prefer traveling and searching for food at night. They are creatures of habit and almost invariably travel from their shelter to their food sources and to the outside over the identical pathways. Perhaps for protection, their pathways ordinarily are in narrow, out-of-the- way locations, such as overhead pipes and beams, or alongside walls, and when rats are running from one location to another, they hug the wall. Rat runs are easily located because dirt and oil from the hair on the rats rub off and blacken the surfaces that they contact.

5. Fourteenth Grade Version

Some interesting and significant factual data about rats which should be remembered when rodent-proofing buildings are: (1) rats can enter crevices as restrictive as 1/2 inch in diameter; (2) rats can climb better vertically, and they can ascend pipes 4 inches in diameter or smaller; (3) rats can jump 3 feet high from flat surfaces, and they can jump 4 feet horizontally; (4) rats can jump 8 feet from an elevated position, and they can fall 50 feet without injuring themselves. Additionally, rats prefer traveling and foraging for food during the nighttime, and because they are creatures of habit, almost invariably travel from their harborage to their food sources and to the outside over the identical pathways. Perhaps for protection, their pathways ordinarily are limited to narrow, out-of-the-way locations, such as overhead pipes and beams, or alongside walls, and when rats are running from one location to another they hug the wall. Rat runs are easily identified because dirt and oil from the hair on the rats rub off and blacken the surfaces that they contact.